Deep Learning on Point Clouds for 3D Object Generation and Classification

Deep Learning on Point Clouds for 3D Object Generation and Classification
(点群深層学習による三次元物体の生成と分類)

Cherdsak KINGKAN

DISSERTATION SUMMARY

Chapter 1 Introduction

As the world around us consists of 3D geometry, 3D understanding enables better human-computer interactions than its 2D counterparts for several applications such as robotics, AR/VR, self-driving vehicle, or medical imaging. 3D data representations have many advantages over its 2D counterpart such as explicit geometry, and illumination invariance, allowing precise and accurate data usage for 3D manipulation, coordination and visualization. 3D data can be represented in several forms, for example, depth image, volumetric grids, polygon mesh, or point cloud. In this dissertation, we are interested in point clouds that can be easily acquired by low-cost 3D sensors for 3D object generation and classification. A point cloud is a collection of points in 3D space. Point clouds provide a simple, and scalable geometric representation suitable for a wide range of applications. For object generation and classification using point clouds, traditional methods employ hand-engineered features extracted from point clouds. Therefore, learning-based approaches are required to improve the performance and computational efficiency.

The recent success of deep learning for images encourages the adaption to geometric data like point clouds. However, the traditional network architectures take as input data with regular structure like those of image pixels, while point clouds are fundamentally irregular. One common method to process point cloud data for deep learning models is to convert raw point cloud data into a volumetric representation, namely 3D voxel grids. However, this approach tends to use a huge amount of memory and also introduce quantization artifacts, making it difficult to capture fine-grained features. This dissertation addresses the aforementioned problems of deep learning on point clouds for 3D object generation and classification tasks.

Chapter 2 Generating Mesh-based Shapes from Learned Latent Spaces of Point Clouds with VAE-GAN

We propose a framework that generates mesh-based objects from its corresponding point clouds using a combination of variational autoencoder (VAE) and generative adversarial network (GAN). The network mainly consists of three parts: (1) the encoder, (2) local and global generators, and (3) the discriminator. Instead of converting point cloud to other representations like voxels before input into the network, our network directly consumes the point cloud and generates the corresponding 3D object. Given point clouds of objects, the encoder encodes local and global geometry structures of point clouds into latent representations. The local and global generators then leverage these latent vectors to generate the implicit surface representations of objects corresponding to those point clouds. Here, the implicit surface representation is Signed Distance Function (SDF), which preserves the inside-outside information of objects. Then we can easily reconstruct polygon mesh surfaces of objects from the SDFs. This could be very helpful in a situation where there is a need of 3D shapes and only point clouds of objects are available. Experiments demonstrate that our network, which makes use of both local and global geometry structure, can correctly generate high-quality mesh-based objects from its corresponding point clouds. We also demonstrate that the size of the latent spaces affects the quality of reconstructed objects.

Chapter 3 Learning Ensembles of Points via Sampling Network for 3D Object Classification

We present a new deep learning architecture for 3D object classification using point cloud data. This network is designed based on the assumption that it is not necessary to process every point in the point cloud to achieve a good classification performance. Instead of learning from all points in the point cloud at once, this model learns to select the most informative points in a feature space for further feature learning in an end-to-end manner. The model selects ensembles of informative points at different layers in order to learn the properties of objects in those spaces. Points are identified as informative points based on the average over the feature dimension of each point. The combination of these ensembles can then be used as shape descriptors for object classification. By learning from a small fraction of points, our model requires a smaller number of training parameters as well as computational cost comparing to other models to achieve the similar performance. We evaluate our model on a standard benchmark dataset like ModelNet40, in which a point cloud of each object consists of 1,024 points. Consequently, our model can achieve competitive classification accuracy with ~35.68% less computational cost comparing to PointNet++ and about a half of training parameters comparing PointNet. We demonstrated that the optimal number of selected points as information points is 512 points in order to achieve the best classification accuracy from our network.

Chapter 4 Point Attention Network for Gesture Recognition Using Point Cloud Data

Understanding human motion has been a popular research topic in recent years, especially in robotics. This is because of increase in situations where humans and robots continue to share spaces. Therefore, there is a need for robots to understand humans' intention through gesture and action recognition. In this chapter, a neural network with novelty attention modules for human gesture recognition from point cloud data is proposed. Our network directly takes point clouds as input, and the attention modules learn to pay attention to only points that contribute to more accurate gesture recognition. The gather and scatter operations are proposed in order to select the most informative points in the feature space for attention mechanism. The attention modules make use of these operations for downsampling and upsampling the number of points in the input point clouds. The input point cloud consists of both spatial and temporal dimensions since the point cloud frames are concatenated as one input sample. This enables the network to learn the spatio-temporal features of the data without explicit modeling of gesture dynamics. We evaluate the performance of our network using a dataset of common Japanese gestures. This dataset consists of 10 classes which are (1) No gesture, (2) Come here, (3) Me, (4) No Thank you, (5) Money, (6) Peace, (7) Not allowed, (8) OK, (9) I am sorry, (10) I got it. Each point cloud contains 2,048 points, and is normalized into [-0.5, 0.5]. As a result, the proposed network achieves state-of-the-art performance with 94.2% recognition accuracy on this dataset, and outperforms the state-of-the-art model by a large margin of ~10%. The architecture design and parameter choices, such as the optimal number of stacked point cloud frames, and the number of attention modules, are also analyzed.

Chapter 5 Learning Hierarchical Probabilistic Latent Spaces of Point Clouds for Mesh-based Shape Reconstruction Using VAE and 3D GAN

In Chapter 5, we leveraged our network in Chapter 3 and redesign the new generator to address the limitations of our model in Chapter 2, i.e., large model size and number of training parameters. We proposed a network for learning a point cloud to 3D object mapping. The encoder directly consumes point clouds as input,

and hierarchically encoded them onto four different probabilistic latent spaces with 32-dimensions. The generator samples latent vectors from those latent spaces, and reconstructs the corresponding signed distance function fields (SDFs) of input point clouds. These SDFs can then be turned into mesh-based shapes. We demonstrate that our network can generate results with high quality surfaces. The performances of the proposed network are also quantitatively and qualitatively compared with the previous state-of-the-art. We can reduce the number of training parameters of our new model, which results in more memory efficiency as well as better reconstructed results comparing to the model in Chapter 2.

Chapter 6 Conclusions

This dissertation focused on mainly two problems, that is, 3D object generation and classification. For 3D object generation, the networks that learn point clouds to 3D object mapping are proposed. Specifically, the networks consist of 3 main components, i.e., an encoder, generators, and a discriminator. The networks take point cloud as an input and generate the corresponding signed distance function (SDF) of an input point cloud. The generated SDF can then be turned into the polygon mesh of an object. Experiments show that the proposed networks are able to learn the point clouds to 3D object mapping, and able to correctly generate the corresponding SDF with high quality surface properties.

For a classification problem, point clouds of both rigid and non-rigid objects are utilized as input. For rigid 3D object classification, the network that learns the ensembles of points in an input point cloud to perform object classification is proposed. The network achieves comparable classification accuracy with less number of training parameters and computational cost comparing to the state-of-the-art network. Also, the network with attention modules to perform gesture recognition using point cloud data is proposed. As a result, the proposed network achieves state-of-the-art performance on the common Japanese gesture dataset.